Comparison between Enriched Travel Data and the Original Survey Data by means of a Model Based Approach

Juliet Nakamya, Hasselt University, Transportation Research Institute
Diepenbeek, Belgium

Elke Moons, Hasselt University, Transportation Research Institute
Diepenbeek, Belgium

Geert Wets, Hasselt University, Transportation Research Institute*
Wetenschapspark 5, bus 6, B-3590 Diepenbeek, Belgium
T: +32 (0) 11 26 91 58
F: +32 (0) 11 26 91 99
E-mail: geert.wets@uhasselt.be

Submission date: March 15, 2008

Submitted for presentation at the 8th International Conference on Survey Methods in Transport, May 2008

*corresponding author

Abstract

This paper focuses on making a comparison between enriched travel data and the original survey data by means of a model based approach. The research reported in this paper represents a continuation of work undertaken in our previous study 'The impact of data integration on some important travel indicators'. The Flemish Household Travel (*FHTS*) data and the Time Use data of the Flemish people (*FTUS*) data were combined in the previous study using the socio-demographic population data as the base data. The *FHTS* data and the resultant combined data are then compared here for the duration of travel per person per day by means of the linear regression model. The explanatory variables exploited include socio-demographic variables. The variable relating to the individual trip rates per person per day was also controlled for and the general analyses are conditioned on individual purpose of travel. The results revealed that the combined data set provided a reduction in the magnitude of the standard errors of the parameter estimates indicating higher precision and thus providing a better basis for purposes of prediction. The larger sample (combined data) is therefore invaluable in prediction of travel demand and offers a better base for simulating travel data. Therefore, integrating data from different sources holds out considerable promise for supplementing existent travel survey data making them more enriched for better prediction.

Keywords: *Data integration, Travel survey data, Time use data*

Introduction

Critical information regarding travel behaviour can be available from several data sources including census data records, sample surveys as well as other administrative data sources. At present however, travel surveys are one of the most important and rich source of the critical information needed for transportation planning and decision making. These surveys are used to collect current data about the demographic, socio-economic, and trip-making characteristics of individuals and households as well as furthering our understanding on travel in relation to the choice, location, and scheduling of daily activities. This facilitates enhancement of travel forecasting methods and improves the ability to forecast changes in daily travel patterns in response to existent social and economic trends as well as new investments in transportation systems and services.

Prompt, quality, and large amounts of data is continuously required from national statistical agencies. Due to the need for informed decision making and policy formulation, the need for these data and information is increasing over the years. In most cases, the provision of large quality data on travel demand, which is related to the socio-demographic and travel characteristics of individuals and households, largely depends on household travel surveys (*HTS*). However, *HTS* are besieged with challenges. They are notoriously expensive and require an appreciable amount of time to plan and implement in spite of the current state of increasingly tight budgets. Even in the face of methodological and technological survey techniques becoming increasingly refined, high unit costs and public resistance are expected to continue to plague future survey endeavors. It is reasonable to believe that even when more sophisticated and recent technologies such as the global positioning system (Murakami and Wagner, 1999; Draijer *et al*., 2000; Murakami *et al*., 2000) and personal digital assistant (Murakami *et al*., 2000; Janssens, *et al*., 2004) are used, the final total cost will only increase. On addition to the just mentioned problems, another, yet big difficulty faced in conducting high-quality travel surveys today is non-participation. Researchers are now getting even more concerned about the high response burden imposed on respondents especially due to the fact that response rates are dropping dramatically. The impact of all these problems on the quality and representativeness of the resulting data is startling.

An increasingly important problem affecting many areas of transport planning, operations and management is the need to combine information from a variety of different data sources in order to provide the best possible estimate of certain parameters of interest.

Combining data from different surveys can be a conceivable option in an effort to lower respondent burden and survey expenses. It is a practical solution that makes use of as much as possible all the information already available in different data sources, that is, to carrying out a statistical integration of data that has already been gathered. While a significant amount of work has been done on data integration (Arellano and Meghir, 1992; Angrist and Krueger, 1992; Winkler, 1995; Lusardi, 1996; D'Orazio *et al*., 2006), most of the research has been performed outside the transportation research community. Nevertheless, to integrate data from different sources in transportation research frequently arises due to several reasons such as overlapping data providing different or conflicting information and the aging of sample survey data and thus the consequent need for updating them. More to this, an important issue in data integration is to measure the quality of the fusion; this is not a trivial problem.

The main aim of this paper is to examine integrated data by making a comparison between enriched travel data and the original survey data using a model based approach. The data available in this study include data from the Flemish Household Travel Survey (*FHTS*) carried out in Belgium in 2000 (Zwerts and Nuyts, 2004), the Flemish Time Use Survey (*FTUS*) also carried out in Flanders in Belgium in 1999 (Glorieux, 2000) and additionally, the Socio-Economic population census (*SEE*) data of 2001 conducted in Belgium (Statistics Belgium, 2001) are also available. The *FHTS* and the *FTUS* data are combined (Nakamya *et al.*, 2007) using the *SEE* as the base data. The *FHTS* data and the resultant combined data are then compared here for the duration of travel per person per day by use of the linear regression model. The explanatory variables exploited include socio-demographic variables. The variable relating to the individual trip rates per person per day was also controlled for and the general analyses conditioned on individual travel purposes.

The rest of this paper is organized as follows. Section 2 describes the surveys that resulted in the data available in this study. In Section 3, the methodology used in this study is laid out. The results of the model estimation and interpretation are then presented and discussed in Section 4 and finally, Section 5 gives the concluding remarks and some directions for further research are presented.

Data

The survey data utilized in this research arise from two surveys: the Flemish Household Travel Survey (*FHTS*) carried out in 2000 (Zwerts and Nuyts, 2004) and the Flemish Time

Use Survey (*FTUS*) carried out also in Flanders, Belgium in 1999 (Glorieux, 2000). Table 1 offers a comparison of the sample design of the *FHTS* and the *FTUS* surveys.

Table 1: A Comparison of the Sample Design of the *FHTS* and the *FTUS* Surveys

	FHTS	*FTUS*
Research population	Flanders	Flanders (incl. Flemings in Brussels)
Age	6 years and above	16-75 years
Sampling-unit	Households	Individuals
Fieldwork	12 months	+- 5 months
N persons	7626	1533
N Households	3027	Not applicable
Sampling	Stratified sample (age of head of household)	Stratified sample (community)
Contacting procedure	By telephone/post or exclusively by post	Introduction letter and 2 face-to-face visits
Research instruments	- Household Questionnaire - Individual Questionnaire - Travel Questionnaire (2 days/ retrospective)	- Individual Questionnaire - Diaries (7 days/ simultaneous)

The *FHTS*, which is the main survey of interest in this study, was conducted among the Flemish citizens. The *FHTS* field work was carried out during a period of 12 months among the Flemish citizens aged 6 years and above. Respondents from a stratified sample of 3,027 households comprising 7,626 persons were asked to fill in an individual questionnaire and also to keep a travel diary for two days. In the travel diary, respondents recorded their travel activities, modes of transport, duration, location, company of others when traveling and search for car parking. The individual questionnaire included socio-demographic variables as well as travel-related variables. Further data was collected from these households by use of household questionnaires. This survey had a response rate of 32% of the households. The second survey, *FTUS,* was carried out by the *Tempus Omnia Revelat* research group of the Free University of Brussels amongst the Flemish citizens. The fieldwork took place between April 15 and October 30, excluding the period between the 15th of July and the 1st of September in 1999. In this survey, 1,533 Flemish people between the ages of 16 and 75 years were requested to record all their activities in a diary for a full week. There were also questions about subsidiary activities, starting and end times, locations, eventual means of

transportation, presence of others, conversation partners during the activity and the motivation to carry out the activity. Regarding the activities, the respondent could make use of a pre-coded list of 154 detailed categories of activities, based on the international time-use study (Szalai, 1972). In addition to the diary registration of *FTUS*, individual questionnaires were also presented to the same sample including socio-demographic variables as well as general indicators on time use and cultural participation. Further more, respondents were asked their opinion about different social issues. A response rate of 28% of the individuals was obtained in this survey.

Data Preparation and Methodology

The data from the two surveys available in this general research were separately cleaned and several variables were adjusted or created in relation to travel. Homogenization (Koelet *et al.,* 2006) of different data sources generally involves a great deal of preliminary effort in practice. At the end of this cascading process, some social demographic variables together with travel related variables were then compatible to each pother in the two surveys. To ensure representativity, the two surveys, which come from the same population, Flanders, were each weighted with respect to the population. In general the weights w_i for class *i* are obtained as:

$$w_i = \frac{P_i}{p_i} \qquad (1)$$

where P_i is the proportion of the *i-th* class in the population and p_i is the proportion of the *i-th* class in the sample. Typical choices of weighting variables are socio-demographic key variables such as geographic indicators or age-gender groups.

Iterative Proportional Fitting (*IPF*) method *(*Bishop *et al.,* 1995) is a well established technique with the theoretical and practical considerations behind the technique clearly examined and reported. This procedure was developed for combining information from two or more sets of data (Bishop *et al.,* 1995). Since in this study socio-demographic population data is fully available however, the internal population frequency cell values for the respective classes of interest are directly utilized to obtain the weights instead of using population marginal values.

Currently policy formulation requires information that is as rich and as timely as possible. The available data may be insufficient and the constraints involved in collecting new

data are often enormous. This provides an urgent driving force to integrate data from different sources. The need to integrate data from different sources thus arises due to different reasons leading to a full range of data combination problems: Data sources may provide both direct and indirect information on some relevant population parameters; the data sources may have varying levels of statistical precision or confidence; data may be overlapping but providing different or conflicting information; the aging of sample survey data and its consequent need for updating; and more so, there may be simply opportunities from new data inflow presenting benefits in updating already existing data.

A number of potential approaches can be deployed to deal with the data integration problem posed above. Integration of data from various sources can be performed by means of three different methodologies: record linkage, merging and statistical matching. The record linkage and merging techniques are considerably different from the statistical matching problem. They are meant to link similar units from two or more different files. Merging requires error-free matching variables, while record linkage is a statistical decision method that can be used when matching-variables are affected by errors. Both techniques require that the sets of units in the two sources overlap. Statistical matching, which is also the technique that was ustilized for the combined data used in this paper, targets providing joint information on variables observed in different sources. It is faced with the problem of integration when the files lack unit identifiers or do not contain the same units.

Data integration, which is achieved through statistical matching is initiated by two or more samples, one usually larger than the other with a negligible overlap of units (such as individuals) in both samples. D'Orazio *et al*., (2006) dealt with the statistical matching problem providing a consistent maximum likelihood estimator of the elements characterizing uncertainty. There are two broad groups of objectives for statistical matching: the *micro* and *macro* objectives. The *micro* approach is obtained when interest is essentially in integrating the database at unit level, and the *macro* approach, when most concern is in the aggregates. Statistical matching methodologies should therefore be chosen according to these two previous objectives.

As discussed by D'Orazio *et al.* (2006), integration of two or more sources of data means the possibility of having joint information on the non-jointly observed variables of the different sources. Our line of research is directed to the micro approach, in which case, the goal is the construction of a *synthetic* file which is *complete*. The file is *synthetic* because it is not a result of direct observation of a set of units in the population of interest, but rather obtained by exploiting information in the source files in some appropriate way. One of our

future aims is to utilize these data as input in micro-simulation models. As such, the creation of this file, which we term as 'a combined data set', is deemed necessary. The file is *complete* in the sense that all the variables of interest, although collected in different sources (*FHTS* and *FTUS*), are contained in it. Using the *FHTS* data containing n_1 persons together with the *FTUS* data, containing n_2 persons, a combined set of data comprising of $(n_1 + n_2)$ units was thus obtained.

The study conducted here represents a continuation of work undertaken by Nakamya, *et. al.,* 2007. Nakamya and colleagues investigated on the impact of data integration on some important travel behavior indicators. Using the same data sets available here, their study conducted the data integration procedure and obtained a combined data set by integrating the two sets of survey data (*FHTS* and *FTUS*) on some socio-demographic characteristics and some common travel characteristics. The combined data set was found to offer a larger and more representative sample of the population, which gives more reliable travel information on the population. The study also provided a set of general guidelines on what practitioners may consider when intending to perform data integration. These include: examining background information on travel data and other data sources; reconciliations of concepts and definitions; re-categorization, re-coding and transformation of variables; and harmonizing time periods of pre-integration data sets.

As mentioned before, an important issue in data integration is to measure the quality of the integration and to further examine these data. To examine the combined data in comparison to the *FTUS* data, a linear regression model is applied separately to each of the dataset. The general linear regression model with $p-1$ predictor variables assumes the form:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{i,p-1} + \varepsilon_i \qquad (2)$$

with

$$E[Y_i] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{i,p-1} \qquad (3)$$

where $\beta_0, \beta_1, ..., \beta_{p-1}$ are parameters, $X_{i1}, X_{i2}, ..., X_{i,p-1}$ are known constants, $i = 1, ..., n$ and ε_i are independent $N(0, \sigma^2)$ that is, normally distributed with zero mean and constant variance σ^2. The general linear model with normal error terms implies that the observations Y_i are independent normal variables, with mean $E\{Y_i\}$ and with constant variance σ^2. More details about linear regression models can be found in Neter *et al.*, (1996).

Activities were frequently divided into two types in the past: work and leisure. This two-way classification has been used in several studies including activity-based trip

generation modeling (Supernak, *et al.*, 1983 and Munshi, 1993). On the other hand, modern consumer theory (Reichman, 1977, and Lane and Lindquist, 1988) typically uses a three-way categorization of activities into: (1) subsistence (income-producing or paid time, such as work); (2) nondiscretionary (obligated, maintenance or compulsory activities, such as eating meals, certain shopping, and child care); and (3) discretionary or leisure activities. Golob, *et al.* (1994), Golob (1998) and Golob and McNally (1997) employ this classification in modeling relationships between activity and travel time. In this study, a classification of the travel purpose attribute closely related to the latter 3-group classification is used. The first group, subsistence, comprises of work, business visits and trips due to following education. The second is the maintenance group which covers all shopping trips, picking/dropping someone and other personal maintenance trips such as obtaining services from doctors, the bank etc. Thirdly, the out-of-home leisure group incorporates trips for visiting purpose, free time sports, culture and relaxation like walking around.

Results and Discussion

Analyses are carried out on the *FHTS* data and the combined data for Flemish respondents aged between 16-75 years. This age range was considered since the *FTUS* data comprise of only respondents between the ages of 16-75 years.

Table 2 shows the distributions of the weighting variables; gender (male, female), age (16-34, 35-54 and 55-75 years of age), marital status (married, divorced, widowed and un-married) and education level (primary school, junior high school, high school and college or university). This is shown for respondents with respect to the socio-economic census data of 2001, the *FHTS*, the *FTUS* and the combined data. Overall, the *FHTS* and the *FTUS* data distributions are very close to the population (socio-economic census) distributions. In the combined sample as well as in the other samples, the majority of the people are married and the least are widowed. It is also noted that most respondents are between 35 and 54 years. Regarding education level, the minority attained at most a certificate of primary education, while most of the respondents have a high school diploma as the highest obtained degree of education. The sample is approximately equally distributed between males and females.

Table 2: Comparison of the Percentage of Respondents by Socio-demographic Factors with Respect to the *FHTS*, *FTUS* and the Combined Data (16-75 Years)

Socio-demographic characteristics	Socio-Economic census 2001 data	*FHTS* data	*FTUS* data	Combined data
Gender				
Male	50	49.80	49.63	49.77
Female	50	50.20	50.37	50.23
Age group				
16-34 years	32	31.79	31.77	31.79
35-54 years	39	39.01	39.80	39.17
55-75 years	29	29.20	28.43	29.05
Marital Status				
Married	62	61.42	61.55	61.44
Divorced	7	7.21	7.20	7.21
Widowed	4	4.46	4.27	4.43
Un-married	27	26.91	26.97	26.92
Education level				
Primary school	18	17.84	15.43	17.35
Junior high school	25	25.42	25.96	25.53
High school	33	32.51	34.36	32.88
College or University	24	24.23	24.25	24.23

The response variable of interest here is the duration of travel per person per day (in minutes). Very often respondents have a tendency of rounding off the durations/time for which they conduct activities. As a result the majority are inclined to record the durations of their travel activities in intervals of 5. However, since there are still some few respondents who record the duration precisely (e.g. they record that they traveled for 9 minutes instead of 10 or 3 minutes instead of 5 minutes), this presents a challenge in modeling the duration due to the nature of the distribution. This was the case with the data in this study. An illustration is given in Figure 1, which shows the distribution of the duration of travel per person per day for the *FHTS* data considering the subsistence purpose of travel (durations greater than 200 minutes are excluded in the plot). The figure shows high peaks at every interval of 5 and sudden drops for other values. The solution was therefore to 'discretize' the entire data set into intervals of 5 so as to obtain a distribution closer to the normal. The resultant distribution in now shown in Figure 2 for the same data. Since the distribution was somewhat still far from normal, the interval-duration of travel was further log-transformed. This variable is henceforth referred to as *logDuration*.

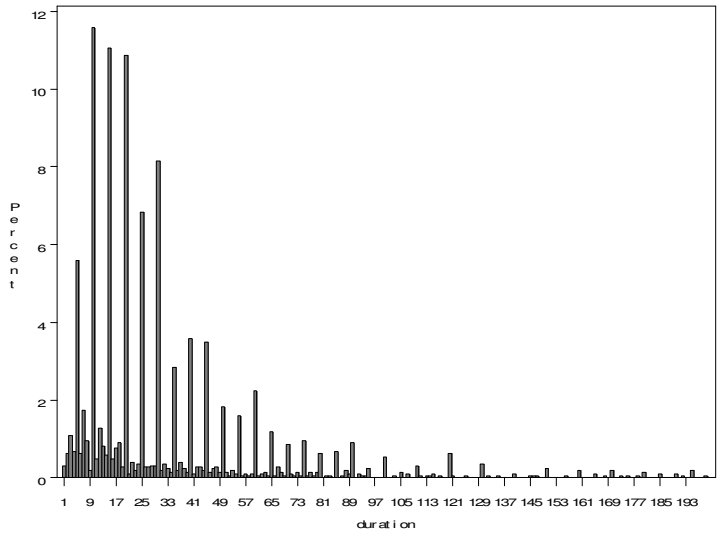

Figure 1: The distribution of the original continuous duration of travel for the FHTS-subsistence purpose

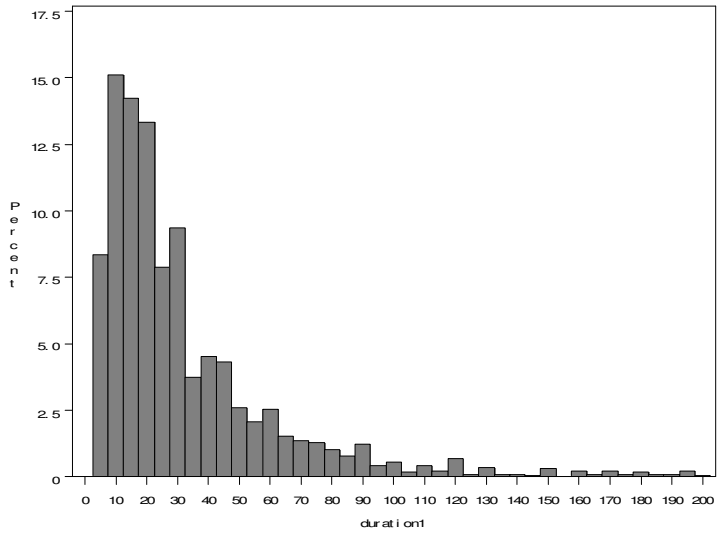

Figure 2: The distribution of the interval duration of travel for the FHTS- subsistence purpose

The linear regression model was separately applied to the FHTS and the combined data to investigate the difference between the fused data and the original survey data. For each of the 3 travel purposes, the full model comprises of the trip rates variable (*TripRate*) and the 4 socio-demographic variables: gender (male or female), age group (*Age*: '16-34', '35-54', '55-75'), marital status (*MS*: married, divorced, widowed , unmarried) and education level (*Edu*: Primary school, junior high school, high school and college/university) together with all possible two-way interactions between the variables. *TripRate* represents the daily

number of trips per person. The full models are reduced through a model reduction procedure which involves comparing more complex models with the reduced ones.

For the subsistence purpose, the model building process resulted into most of the interaction effects together with the age group main effect being statistically insignificant. These were therefore dropped from the model. Table 3 shows the parameter estimates from the final linear regression model of *logDuration* with respect to the *FHTS* and the combined data. The interaction effects that were found to be significant include trip rates with education level and trip rates with marital status.

Table 3: Results of the Linear Regression Model for the Log-transformed Duration of Travel per Person per Day with respect to the *FHTS* and the Combined Data (Subsistence Travel Purpose)

	FHTS data (n=2257, R^2=0. 1720)			Combined data (n=2831, R^2=0.1732)	
Variable	Parameter estimate (standard error)	p-value		Parameter estimate (standard error)	p-value
Intercept	2.868 (0.06595)	<.0001		2.853 (0.06016)	<0.0001
TripRate	0.208 (0.03676)	<.0001		0.240 (0.03420)	<0.0001
Gender-male	0.145 (0.03448)	<.0001		0.164 (0.03045)	<0.0001
MS-married	-0.123 (0.06505)	0.0596		-0.075 (0.05926)	0.2073
MS-divorced	-0.167 (0.15825)	0.2927		-0.153 (0.14326)	0.2861
MS-widowed	-1.233 (0.46535)	0.0081		-1.028 (0.43253)	0.0176
Edu-primary	-0.756 (0.19441)	0.0001		-0.774 (0.17587)	<0.0001
Edu-junior high	-0.316 (0.08479)	0.0002		-0.385 (0.07746)	<0.0001
Edu-high	-0.210 (0.07235)	0.0037		-0.240 (0.06524)	0.0002
*TripRate** *Edu*-primary	0.375(0.15804)	0.0178		0.369 (0.14204)	0.0094
*TripRate** *Edu*- junior high	0.110 (0.05308)	0.0382		0.141 (0.04941)	0.0044
*TripRate** *Edu*- high	0.086 (0.04318)	0.0456		0.093 (0.03933)	0.0184
*TripRate** *MS*-married	0.104 (0.04038)	0.0101		0.065 (0.03723)	0.0806
*TripRate** *MS*-divorced	0.224 (0.11078)	0.0431		0.199 (0.09998)	0.0466
*TripRate** *MS*-widowed	0.725 (0.38376)	0.0590		0.649 (0.34811)	0.0624

The results corresponding to the *FHTS* and the combined data are quite close. In general, the parameter estimates are consistent in direction and magnitude. The estimated regression function indicates that the mean log of duration of travel is expected to increase by 0.145 (or 0.164 for the combined data) for males relative to females with other factors held constant. Also in the regression model, the effect of trip rates per person per day differs for different levels of education and in the same way, the effect of education on *logDuration* differs for different levels of trip rates when other factors are held constant. A similar

interpretation applies due to the interaction effect of trip rates with marital status. Overall, the combined data set provides a reduction in the magnitude of the standard errors of the parameter estimates thus providing a better basis for purposes of prediction. The dataset also gives a slightly higher R-square for the model implying slightly higher explained variability in the response. On the whole, however, the R-square values are all low, which is a result also frequently observed in sociological studies.

For the analyses conditional on the maintenance travel purpose, all interaction effects together with the main effect education are not found to be statistically significant and were therefore dropped from the model. Table 4 shows the results of the final linear regression model for *logDuration.* Results are displayed for parameter estimates from the final model with respect to the *FHTS* and the combined data.

Table 4: Results of the Linear Regression Model for the Log-transformed Duration of Travel per Person per Day with respect to the *FHTS* and the Combined Data (Maintenance Travel Purpose)

	FHTS data (n=2539, R^2=0. 2367)			Combined data (n=3162, R^2=0.2143)	
variable	Parameter estimate (standard error)	p-value		Parameter estimate (standard error)	p-value
Intercept	2.426 (0.05723)	<.0001		2.37640 (0.05670)	<0.0001
TripRate	0.390 (0.01437)	<.0001		0.41387 (0.01445)	<0.0001
Gender-male	0.120 (0.03099)	0.0001		0.07907 (0.03048)	0.0095
MS-married	-0.116 (0.04384)	0.0084		-0.15074 (0.04326)	0.0005
MS-divorced	-0.174 (0.06923)	0.0119		-0.20718 (0.06880)	0.0026
MS-widowed	-0.027 (0.09262)	0.7686		-0.082 (0.09160)	0.3713
Age(16-34)	-0.226 (0.04623)	<.0001		-0.145 (0.04571)	0.0015
Age(35-54)	-0.132 (0.03851)	0.0006		-0.060 (0.03811)	0.1161

The final model, which is again the same model obtained from both datasets, comprises of the trip rates variable, gender, marital status and the age group variable. In general, the results corresponding to the *FHTS* and the combined data are quite close with the parameter estimates consistent in direction and magnitude. The estimates of standard errors are roughly smaller for the combined data set indicating higher precision. The duration of travel per person per day generally increases with increase in trip rates. Males are observed to have higher maintenance travel duration as compared to females. All types of marital status are seen to travel less relative to the unmarried although the difference between the unmarried and the widowed is not statistically significant. Lower age groups spend lower duration travel

on maintenance as compared to the 55-75 age group. However, the difference between the '35-54' and the '55-75' age group is not statistically significant.

The final analyses were conditioned on the out-of-home leisure travel purpose. The final model that was obtained here incorporated only two variables relating to trip rates and age groups.

Table 3: Results of the Linear Regression Model for the Log-transformed Duration of Travel per Person per Day with respect to the *FHTS* and the Combined Data (Out-of-home Leisure Travel Purpose)

	FHTS data (n=1976, R^2=0.2398)			Combined data (n=2660, R^2=0.2403)	
variable	Parameter estimate (standard error)	p-value		Parameter estimate (standard error)	p-value
Intercept	2.28619 (0.05023)	<.0001		2.37772 (0.04590)	<0.0001
TripRate	0.60818 (0.02497)	<.0001		0.61890 (0.02192)	<0.0001
Age(16-34)	-0.23648 (0.04870)	<.0001		-0.30000 (0.04512)	<0.0001
Age(35-54)	-0.10306 (0.04879)	0.0348		-0.12569 (0.04529)	0.0026

Higher precision is obtained with the combined data set as compared with the *FHTS* data and a slightly higher R-square is obtained. Just as observed in the results before, the parameter estimates here are consistent in direction and magnitude. Higher trip rates per person per day generally correspond to higher out-of-home leisure durations. The '55-75'age group spends significantly higher duration of travel on out-of-home leisure as compared to other age groups. This could be because most of these people are pensioners and thus have more time to spend on such activities. The '16-34' age group is observed to be traveling the least for out-of-home leisure.

Conclusions and Future Research

This paper utilizes combined data from a household travel survey and a time use survey to make a comparison between the resultant enriched travel data and the original survey data using a model based approach. The combined travel data provides a larger and more representative sample of the population, which gives more reliable travel information on the population (Nakamya *et al.*, 2007).

The analyses in this study involved use of the linear regression model to compare the combined data to the original Flemish household travel survey data. The main response

variable was the duration of travel per person per day and mainly socio-demographic variables serve as the explanatory variables. The variable relating to the individual trip rates per person per day was also controlled for and the general analyses were conditioned on the individual travel purpose.

The results presented here generally show that the combined data provide parameter estimates that are consistent in direction and magnitude for all purposes of travel as compared to the original travel survey data for the estimated regression function. Also overall, the combined data set provided a reduction in the magnitude of the standard errors of the parameter estimates indicating higher precision and thus providing a better basis for purposes of prediction. The larger sample (combined data) is therefore invaluable in prediction of travel demand and offers a better base for simulating travel data. Consequently, integrating data from different sources holds out considerable promise for supplementing existent travel survey data making them more enriched for better prediction.

Future research will be directed towards simulation of travel data, to validation as well as to investigation on further improvements involving local data updates. It is expected that this approach will enable Flanders and other regions or countries to develop a synthetic local travel data set and estimate travel-demand models at a proportion of the cost of conducting a traditional household travel survey.

References

Angrist, J. D. and Krueger, A. B. (1992). 'The Effect of Age at School on Entry on Educational Attainment: An Application of Instrumental variables with Moments from Two Samples'. *Journal of the American Statistical Association*. **87 (**418**)**.

Arellano, M. and Meghir, C. (1992**). '**Female Labor Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets'. *Review of Economic Studies*, **59(**3), pp. 537-559.

Bishop, M. M., Fienberg, S. E. and Holland, P. W. (1995). *Discrete Multivariate Analysis: Theory and Practice.* The MIT, England.

D'Orazio, M., Di Zio, M. and Scanu**,** M. (2006) *Statistical Matching: Theory and Practice*. John Wiley and Sons, Inc., New York.

Draijer, G., Kalfs, N. and Perdok, J. (2000) 'Global positioning system as data collection method for travel research' *Transportation Research Record,* **1719,** 147-153.

Glorieux, I., Koelet, S. and Moens, M. (2000) 'Technisch verslag bij de tijdsbudgetenquête TOR'99: Veldwerk en responsanalyse. (tor2000/43)' Vrije Universiteit Brussel, Belgium (in Dutch).

Golob, T.F. (1998). A model of household demand for activity participation and mobility. In T. Gärling, T. Laitila and K. Westin, eds*., Theoretical Foundations of Travel Choice modeling*, 365-398. Pergamon, Oxford.

Golob, T.F. R. Kitamura and C. Lula (1994). "Modeling the Effects of Commuting Time on Activity Duration and Non-Work Travel." Presented at Annual Meeting of Transportation Research Board, Washington, DC, January.

Golob, T.F. and M.G. McNally (1997). A model of household interactions in activity participation and the derived demand for travel. *Transportation Research*, **31B,** 177-194.

Greaves, S. P. (2006) 'Simulating Household Travel Survey Data' University of Sydney.

Janssens, D., Wets, G., Brijs, T. and Vanhoof, K. (2004) 'Simulating Activity Diary Data by Means of Sequential Probability Information: Development and Evaluation of an Initial Framework' Paper Presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, DC.

Kulkarni, A.A. and McNally, M.G. (2001), 'A Microsimulation of Daily Activity Patterns'

Paper Presented at the 80th Annual Meeting of the Transportation Research Board, Washington DC.

Lane, P. M. and J.D. Lindquist (1988). Definitions for the fourth dimension: A proposed time classification system. In K.D. Bahn, ed., *Developments of Marketing Science II.* Academy of Marketing Science, Blacksburg, VA.

Lusardi, A. (1996) 'Permanent Income, Current Income, and Consumption: Evidence from Two Panel Data Sets' *Journal of Business and Economic Statistics*. **14** (1).

Munshi, K. (1993). Urban passenger demand estimation: A household activity approach. *Transportation Research*, **27A**, 423-432.

Murakami, E. and Wagner, D. P. (1999) 'Can using global positioning system (GPS) improve trip reporting?' *Transportation Research C*, 7(2/3), 149.165.

Murakami, E., Wagner, D. P. and Neumeister, D. M. (2000) 'Using global positioning systems and personal digital assistants for personal travel surveys in the United States' Transport Surveys: Raising the Standards, Transportation Research Circular, E-008, TRB, National Research Council, Washington, D.C., III-B/1-21.

Nakamya, J., Moons, E. and Wets, G. (2007) 'The Impact of Data Integration on Some Important Travel Behavior Indicators' *Transportation Research Record*, **1993**, 89-94.

Neter, J., Kutner, M. H., Nachtsheim, C.J., Wasserman, W. (1996) *Applied Linear Statistical Models. 4th ed. The McGraw-Hill Companies, Inc., USA.*

Reichman, S. (1977). Instrumental and life style aspects of urban travel behavior. *Transportation research Record, **649**,* 38-42.

Rubin, R. B. (1987) *Multiple Imputation for Nonresponse in Surveys,* John Wiley & Sons, New York.

Statistics Belgium. http://www.statbel.fgov.be/ Accessed March, 2008.

Supernak, J., Talvitie, A. and DeJohn, A. (1983) 'Person-Category Trip-Generation Model' *Transportation Research Record,* **944**, 74-83.

Szalai, A. (1972) 'The Uses of Time: Daily Activities of Urban and Suburban Populations in Twelve Countries' The Hague, Mouton.

Winkler, W. E. (1995) "Matching and Record Linkage," in B. G. Cox *et al*. (ed.) *Business Survey Methods*, John Wiley and Sons, Inc., New York, pp. 355-384.

Zwerts, E. and Nuyts, E. (2004) 'Onderzoek Verplaatsingsgedrag Vlaanderen 2 (D/2004/3241/016)' Provinciale Hogeschool Limburg, Diepenbeek (in Dutch).